Modern Theory of 2nd-Order Methods

Lecture 3: Universal 2nd-order methods

Yurii Nesterov (CORE/INMA, UCLouvain)

Minicourse: August 16-20, 2021 (Zinal)

Contents

Problem formulation

Let $B = B^* : \mathbb{E} \to \mathbb{E}^*$, and $B \succ 0$. Denote

$$\|x\| = \langle Bx, x\rangle^{1/2},\ x \in \mathbb{E}, \ \text{ and } \|s\|_* = \langle s, B^{-1}s\rangle^{1/2},\ s \in \mathbb{E}^*.$$

For $A : \mathbb{E} \to \mathbb{E}^*$, we can define the matrix norm:

$$\|A\| = \max_{x\in\mathbb{E}}\{\|Ax\|_*,\ \|x\| \le 1\} \quad \Leftrightarrow \quad \langle Ax, x\rangle \le \|A\| \cdot \|x\|^2 \quad \forall x \in \mathbb{E}.$$

Consider the problem of *Composite Minimization*

$$\boxed{(A_3) \quad \min_{x\in\mathbb{E}} \{F(x) = f(x) + \Psi(x)\}}$$

where

- f is a smooth closed function,
- Ψ is a *simple* closed convex function with $\mathrm{dom}\,\Psi \neq \emptyset$.
- $\mathrm{dom}\, f \supset \mathrm{dom}\,\Psi$.

Example: $\Psi(x) = \mathrm{Ind}\ Q$, where Q is a closed convex set.

Main assumption

Define a system of Hölder constants:

$$H_f(\nu) = \sup_{\substack{x,y \in \mathbb{E} \\ x \neq y}} \frac{\|f''(x) - f''(y)\|}{\|x-y\|^{\nu}}$$

with $\nu \in [0,1]$.

Assumption 1. $H_f(\nu) < +\infty$ at least for one $\nu \in [0,1]$.

Lemma. Constant $H_f(\cdot)$ is log-convex functions of ν.

Proof. Indeed,

$$\ln H_f(\nu) = \sup_{\substack{x,y \in \mathbb{E} \\ x \neq y}} \Big\{ \ln \|f''(x) - f''(y)\| - \nu \ln \|x - y\| \Big\}.$$

This is a convex function in ν. □

Thus, if $H_f(0) < +\infty$ and $H_f(1) < +\infty$, then

$$\boxed{H_f(\nu) \leq H_f^{(1-\nu)}(0) H_f^{\nu}(1)}$$

for all $\nu \in [0,1]$.

Examples

1. If $H_f(1) < \infty$, we have Lipschitz condition for Hessians:

$$\|f''(x) - f''(y)\| \leq H_f(1)\|x - y\| \quad \forall x, y \in \mathbb{E}.$$

2. If $H_f(0) < \infty$, we have functions with *bounded variation* of Hessian:

$$\|f''(x) - f''(y)\| \leq H_f(0) \quad \forall x, y \in \mathbb{E}.$$

This is true for $f(x) = \sum_{k=1}^{m} (\langle a_i, x\rangle - b_i)_+^2$, where $(\tau)_+ = \max\{\tau, 0\}$.

This function has *discontinuous* Hessian.

NB: Complexity of problem (A_3) is not changing after addition of *arbitrary* quadratic function.

Hence, the usual *condition number* does not work.

Main inequalities

Theorem. For all $x, y \in \mathrm{dom}\, f$ we have

$$(B_3) \quad \begin{cases} \left| f(y) - f(x) - \langle f'(x), y - x \rangle - \frac{1}{2} \langle f''(x)(y-x), y - x \rangle \right| \\ \qquad \leq \frac{H_f(\nu) \|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)}, \end{cases}$$

$$(C_3) \quad \begin{cases} \| f'(y) - f'(x) - f''(x)(y-x) \|_* \\ \qquad \leq \frac{H_f(\nu)}{1+\nu} \|y - x\|^{1+\nu}. \end{cases}$$

Proof: by integration. □

Regularized Newton Step

Denote

$$\begin{aligned} Q(x;y) &= f(x) + \langle f'(x), y-x\rangle + \tfrac{1}{2}\langle f''(x)(y-x), y-x\rangle, \\ M_{\nu,H}(x;y) &= Q(x;y) + \tfrac{H}{(1+\nu)(2+\nu)}\|y-x\|^{2+\nu} + \Psi(y). \end{aligned}$$

Main property:

$$\text{if } H \geq H_f(\nu), \text{ then } \boxed{F(y) \leq M_{\nu,H}(x;y)}$$

Regularized step

$$\text{Define } \boxed{T_{\nu,H}(x) = \arg\min_{y\in\mathbb{E}} M_{\nu,H}(x;y)}$$

It is uniquely defined.

Assumption 2. $T_{\nu,H}(x)$ is easily computable.

Main property

Optimality condition

For $T = T_{\nu,H}(x)$, and all $y \in \mathrm{dom}\,\Psi$, we have

$$\langle f'(x) + f''(x)(T - x) + \tfrac{H}{1+\nu}\|x - T\|^{\nu} B(T - x), y - T\rangle$$

$$+\Psi(y) \geq \Psi(T).$$

Denote

$$\Psi'(T) = -\Big(f'(x) + f''(x)(T - x) + \tfrac{H}{1+\nu}\|x - T\|^{\nu} B(T - x)\Big).$$

$$\text{Then } \boxed{\Psi'(T) \in \partial\Psi(T)}$$

Efficiency of Regularized Step

For $T = T_{\nu,H}(x)$ denote $\boxed{F'(T) = f'(T) + \Psi'(T) \in \partial F(T)}$

Lemma 1. If $H \geq (1+\nu)H_f(\nu)$, then

$$F(x) - F(T) \geq \langle F'(T), x - T \rangle \geq \left(\frac{1}{8H}\right)^{\frac{1}{1+\nu}} \|F'(T)\|_*^{\frac{2+\nu}{1+\nu}}.$$

NB. We can ensure $F'(x) \to 0$ for nondifferentiable function $F(\cdot)$.

Hence, for sharp minimum we have *finite termination*.

Simple Method: $\boxed{x_{k+1} = T_{\nu,(1+\nu)H_f(\nu)}(x_k)}$ Then

$$F(x_k) - F(x_{k+1}) \geq \left(\frac{1}{8H}\right)^{\frac{1}{1+\nu}} \left(\frac{F(x_{k+1}) - F^*}{D}\right)^{\frac{2+\nu}{1+\nu}},$$

where $D = \mathrm{Diam}\,\{x : F(x) \leq F(x_0)\}$.

Convergence: $\boxed{F(x_k) - F^* \leq O\left(\frac{1}{k^{1+\nu}}\right)}$ **Complexity:** $O\left(\frac{1}{\epsilon^{\frac{1}{1+\nu}}}\right)$.

Accelerated method with known ν

1. Let $x_0 \in \mathrm{dom}\,\Psi$, $M_0 > 0$. Set $A_0 = 0$ and $\boxed{\psi_0(x) = \frac{1}{2+\nu}\|x - x_0\|^{2+\nu}}$

2. For $t \geq 0$ iterate: **a)** Find $v_t = \arg\min_{x \in \mathbb{E}} \psi_t(x)$.

b) Find $i_t \geq 0$ such that for a_{t+1} defined by $\boxed{a_{t+1}^{2+\nu} = \frac{(A_t + a_{t+1})^{1+\nu}}{16 M_t \cdot 2^{i_t}}}$ and

$$A_{t+1} = A_t + a_{t+1}, \;\; \alpha_t = \frac{a_{t+1}}{A_{t+1}}, \;\; y_t = (1-\alpha_t)x_t + \alpha_t v_t,$$

the point $\boxed{x_{t+1} = T_{\nu, M_t \cdot 2^{i_t}}(y_t)}$ satisfies inequality

$$\langle F'(x_{t+1}), y_t - x_{t+1} \rangle \geq \left(\frac{1}{8 M_t \cdot 2^{i_t}} \right)^{\frac{1}{1+\nu}} \|F'(x_{t+1})\|_*^{\frac{2+\nu}{1+\nu}}.$$

c) Define $M_{t+1} = 2^{i_t - 1} M_t$ and

$$\psi_{t+1}(x) = \psi_t(x) + a_{t+1}[f(x_{t+1}) + \langle f'(x_{t+1}), x - x_{t+1} \rangle + \Psi(x)].$$

Theorem 1. $\boxed{F(x_t) - F^* \leq \frac{16 \gamma H_f(\nu)(4+2\nu)^{1+\nu} \|x_0 - x^*\|^{2+\nu}}{(t-1)^{2+\nu}}}$

Complexity: $O\left(\frac{1}{\epsilon^{\frac{1}{2+\nu}}} \right)$, as compared with $O\left(\frac{1}{\epsilon^{\frac{1}{1+\nu}}} \right)$.

Efficiency of the universal step

We need to estimate from above

$$Q(x;y) + \frac{H_f(\nu)\|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)} + \Psi(y) \quad \text{by} \quad Q(x;y) + \frac{H\|y-x\|^3}{6} + \Psi(y).$$

When this can be done? *We need big steps!*

Lemma 2. Let $x_+ = T_{1,H}(x)$. If $\|F'(x_+)\|_* \geq \delta$ and

$$H \geq \left[\frac{CH_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}} \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} \quad \text{with } C \geq 6,$$

then $\|x_+ - x\|^{1-\nu} \geq \frac{CH_f(\nu)}{(1+\nu)(2+\nu)H}$. Hence, $\frac{H_f(\nu)\|x_+-x\|^{2+\nu}}{(1+\nu)(2+\nu)} \leq \frac{H\|x_+-x\|^3}{C}$.

Lemma 3. If $\|F'(x_+)\|_* \geq \delta$ and $H \geq \left[\frac{12H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}} \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}}$,

$$\text{then } \boxed{\langle F'(x_+), x - x_+\rangle \geq \sqrt{\frac{4}{3H}}\|F'(x_+)\|_*^{\frac{3}{2}}}$$

Universal accelerated scheme

Let $\boxed{\theta_\nu(\epsilon) = \left[\frac{12 H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}} \left(\frac{D}{\epsilon}\right)^{\frac{1-\nu}{1+\nu}}}$

Choose $x_0 \in \mathbb{E}$ and $H_0 \leq \inf\limits_{0 \leq \nu \leq 1} \theta_\nu(\epsilon)$.

Set $\psi_0(x) = \frac{1}{3}\|x - x_0\|^3$, and $A_0 = 0$.

k-**th iteration** $(k \geq 0)$ **a)** Compute $v_k = \arg\min\limits_{x \in \mathbb{E}} \psi_k(x)$.

b) Find $i_k \geq 0$ such that for $a_{k,i_k}^3 = \frac{3(A_k + a_{k,i_k})^2}{2(2^{i_k} H_k)}$ used in definitions

$\alpha_{k,i_k} = \frac{a_{k,i_k}}{A_k + a_{k,i_k}}$, $y_{k,i_k} = (1 - \alpha_{k,i_k}) x_k + \alpha_{k,i_k} v_k$, and $x_{k+1,i_k} = T_{1,2^{i_k} H_k}(y_{k,i_k})$

we have $\langle F'(x_{k+1,i_k}), y_{k,i_k} - x_{k+1,i_k} \rangle \geq \left(\frac{4}{3(2^{i_k} H_k)}\right)^{\frac{1}{2}} \|F'(x_{k+1,i_k})\|_*^{\frac{3}{2}}$.

c) Set $x_{k+1} = x_{k+1,i_k}$, $A_{k+1} = A_k + a_{k,i_k}$, $H_{k+1} = 2^{i_k - 1} H_k$, and

$\psi_{k+1}(x) = \psi_k(x) + a_{k,i_k}[f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle + \Psi(x)]$.

Convergence results

Theorem 2. Assume $H_f(\nu) < +\infty$ for some $\nu \in [0, 1]$.

And let $\boxed{F(x) - F^* \geq \epsilon}$ at all test points.

Then $H_k \leq 2\theta_\nu(\epsilon)$ for all $k \geq 0$. Moreover, for all $k \geq 2$, we have

$$\boxed{F(x_k) - F^* \leq \frac{48\theta_\nu(\epsilon)D^3}{(k-1)^3}}$$

Complexity result

$$\boxed{k \leq O\left(\left(\frac{1}{\epsilon}\right)^{\frac{2}{3(1+\nu)}}\right)}$$

Calls of oracle: $\leq 2k + \log_2 \theta_\nu(\epsilon)$.

Hint: $H_{k+1} = 2^{i_k - 1} H_k$.

Conclusion

1. We managed to accelerate Regularized Newton Method by aggregating the linear model of the objective function.

2. The complexity results are as follows:

	Universal	Known ν
One-step scheme	$\left(\frac{1}{\epsilon}\right)^{\frac{1}{1+\nu}}$	$\left(\frac{1}{\epsilon}\right)^{\frac{1}{1+\nu}}$
Accelerated scheme	$\left(\frac{1}{\epsilon}\right)^{\frac{2}{3(1+\nu)}}$	$\left(\frac{1}{\epsilon}\right)^{\frac{1}{2+\nu}}$

3. For $\nu = 1$, Universal Scheme is perfect.

4. For $\nu = 0$, we have an extra factor $\left(\frac{1}{\epsilon}\right)^{\frac{1}{6}}$. Can we do better?